REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
18-01-2014	Final Report	29-Apr-2009 - 31-Dec-2013

4. TITLE AND SUBTITLE
Final report

5a. CONTRACT NUMBER
W911NF-09-1-0205

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Wei-Yin Loh

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of Wisconsin - Madison
Suite 6401
21 N Park St
Madison, WI 53715 -1218

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
54050-MA.37

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
Classification and regression tree methodology is an important and essential tool in statistics and machine learning. This research accomplished several improvements and advancements in the area and implemented them in the GUIDE computer software. The major contributions are (i) a new technique to deal with missing data values that allows all the information, including whether or not an observation is missing, to be used for tree construction and prediction, (ii) a new method of scoring the importance of variables that can be used to objectively reduce the number of variables for prediction modeling, (iii) a new approach to building regression models for data with

15. SUBJECT TERMS
classification, regression, tree-structured models, missing values

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Wei Loh
UU	UU	UU	UU		19b. TELEPHONE NUMBER 608-262-7790

Report Title

Final report

ABSTRACT

Classification and regression tree methodology is an important and essential tool in statistics and machine learning. This research accomplished several improvements and advancements in the area and implemented them in the GUIDE computer software. The major contributions are (i) a new technique to deal with missing data values that allows all the information, including whether or not an observation is missing, to be used for tree construction and prediction, (ii) a new method of scoring the importance of variables that can be used to objectively reduce the number of variables for prediction modeling, (iii) a new approach to building regression models for data with multidimensional or longitudinal response variables that does not require any model assumptions, and (iv) several new techniques for identifying subgroups of the data for enhanced differential treatment effects.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received		Paper
07/28/2011	8.00	Wei-Yin Loh. Classification and regression trees, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, (01 2011): 0. doi: 10.1002/widm.8
07/28/2011	12.00	M. E. Piper, J. W. Cook, T. R. Schlam, D. E. Jorenby, S. S. Smith, D. M. Bolt, W.-Y. Loh. Gender, race, and education differences in abstinence rates among participants in two randomized smoking cessation trials, Nicotine & Tobacco Research, (05 2010): 0. doi: 10.1093/ntr/ntq067
08/09/2010	6.00	A. B. Kara, A. J. Wallcraft, H. E. Hurlburt, W.-Y. Loh. Which surface atmospheric variable drives the seasonal cycle of sea surface temperature over the global ocean?, Journal of Geophysical Research: Atmospheres, (03 2009): . doi:
08/09/2010	2.00	Wei-Yin Loh. Improving the precision of classification trees, Annals of Applied Statistics, (01 2009): . doi:
08/09/2010	7.00	Wei-Yin Loh, Wei Zheng. On bootstrap tests of hypotheses, , (01 2009): . doi:
08/11/2011	19.00	Chanyoung Lee, Bin Ran, Fan Yang, Wei-Yin Loh. A Hybrid Tree Approach to Modeling Alternate Route Choice Behavior With Online Information, Journal of Intelligent Transportation Systems, (11 2010): 209. doi:
08/14/2012	25.00	Mounir El Asmar, Wafik Boulos Lotfallah, Wei-Yin Loh, Awad S. Hanna. Uncertainty Reduction in Multi-Evaluator Decision Making, Journal of Computing in Civil Engineering, (01 2012): 0. doi: 10.1061/(ASCE)CP.1943-5487.0000119
08/14/2012	27.00	Wei-Yin Loh, Megan E. Piper, Tanya R. Schlam, Michael C. Fiore, Stevens S. Smith, Douglas E. Jorenby, Jessica W. Cook, Daniel M. Bolt, Timothy B. Baker. Should all Smokers Use Combination Smoking Cessation Pharmacotherapy? Using Novel Analytic Methods to Detect Differential Treatment Effects Over 8 Weeks of Pharmacotherapy, Nicotine & Tobacco Research, (02 2012): 131. doi:
08/14/2012	26.00	Wei-Yin Loh. Variable Selection for Classification and Regression in Large p, Small n Problems, Springer Lecture Notes in Statistics, (01 2012): 133. doi:
08/14/2012	23.00	Timothy B. Baker, Megan E. Piper, Tanya R. Schlam, Jessica W. Cook, Stevens S. Smith, Wei-Yin Loh, Daniel Bolt. Are Tobacco Dependence and Withdrawal Related Amongst Heavy Smokers? Relevance to Conceptualizations of Dependence, Journal of Abnormal, (12 2012): 0. doi:
08/14/2012	22.00	Megan E. Piper, Tanya R. Schlam, Jessica W. Cook, Megan A. Sheffer, Stevens S. Smith, Wei-Yin Loh, Daniel M. Bolt, Su-Young Kim, Jesse T. Kaye, Kathryn R. Hefner, Timothy B. Baker. Tobacco withdrawal components and their relations with cessation success, Psychopharmacology, (03 2011): 569. doi: 10.1007/s00213-011-2250-3
08/14/2012	21.00	Megan E. Piper, Wei-Yin Loh, Stevens S. Smith, Sandra J. Japuntich, Timothy B. Baker. Using Decision Tree Analysis to Identify Risk Factors for Relapse to Smoking, Substance Use & Misuse, (02 2011): 492. doi: 10.3109/10826081003682222

09/12/2013 34.00 Wei-Yin Loh, Wei Zheng. Regression trees for longitudinal and multiresponse data, The Annals of Applied Statistics, (03 2013): 495. doi: 10.1214/12-AOAS596

09/12/2013 35.00 Mounir El Asmar, Awad S. Hanna, Wei-Yin Loh. Quantifying Performance for the Integrated Project Delivery System as Compared to Established Delivery Systems, Journal of Construction Engineering and Management, (06 2013): 0. doi: 10.1061/(ASCE)CO.1943-7862.0000744

10/22/2013 33.00 Wafik Boulos Lotfallah, Wei-Yin Loh, Awad S. Hanna, Mounir El Asmar. Reducing Bias and Uncertainty in Multievaluator Multicriterion Decision Making, Journal of Computing in Civil Engineering, (03 2013): 167. doi: 10.1061/(ASCE)CP.1943-5487.0000206

TOTAL: **15**

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in non peer-reviewed journals:

(c) Presentations

Number of Presentations: 0.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

TOTAL:

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received	Paper

TOTAL:

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received		Paper
07/28/2011	14.00	Wei-Yin Loh. Variable Selection for Classification and Regression in Large p, Small n Problems, Springer Lecture Notes in Statistics (07 2011)
07/28/2011	17.00	Megan E. Piper , Wei-Yin Loh, Stevens S. Smith, Sandra J. Japuntich, Timothy B. Baker. Using decision tree analysis to identify risk factors for relapse to smoking, Substance Use and Misuse (07 2011)
07/28/2011	13.00	Wafik Boulos Lotfallah, Mounir El Asmar, Wei-Yin Loh, Awad S. Hanna. Uncertainty reduction in multi-evaluator decision making, ASCE Journal of Computing in Civil Engineering (07 2011)
07/28/2011	16.00	Megan E. Piper, Tanya R. Schlam, Jessica W. Cook, Megan A. Sheffer, Stevens S. Smith, Wei-Yin Loh, Daniel M. Bolt, Su-Young Kim, Jesse T. Kaye, Kathryn R. Hefner, Timothy B. Baker. Tobacco withdrawal components and their relations with cessation success, Psychopharmacology (07 2011)
08/09/2010	5.00	M. E. Piper, J. W. Cook, T. R. Schlam, D. E. Jorenby, S. S. Smith, D. M. Bolt, W.-Y. Loh. Gender, race, and education differences in abstinence rates among participants in two randomized smoking cessation trials, Nicotine and Tobacco Research (08 2010)
08/09/2010	4.00	Wei-Yin Loh. Classification and regression trees, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery (01 2009)
08/09/2010	3.00	Wei-Yin Loh. Tree-structured classifier, Wiley Interdisciplinary Reviews: Computational Statistics (01 2009)
08/09/2010	1.00	Chanyoung Lee, Bin Ran, Fan Yang, Wei-Yin Loh. A hybrid tree approach to modeling alternate route choice behavior with online information, Journal of Intelligent Transportation Systems (08 2010)
08/14/2012	24.00	Wafik Boulos Lotfallah, Wei-Yin Loh, Awad S. Hanna, Mounir El Asmar. Reducing Bias and Uncertainty in Multi-Evaluator Multi-Criterion Decision Making, Journal of Computing in Civil Engineering (12 2012)
08/14/2012	29.00	Wei-Yin Loh. Fifty Years of Classification and Regression Trees, International Statistical Review (02 2012)
08/14/2012	28.00	Wei Zheng, Wei-Yin Loh. Regression Trees for Longitudinal and Multiresponse Data, Annals of Applied Statistics (08 2012)
09/12/2013	36.00	Wei-Yin Loh. Fifty years of classification and regression trees, International Statistical Review ( )

TOTAL: 12

Number of Manuscripts:

Books

Received Paper

TOTAL:

Patents Submitted

Patents Awarded

Awards

Graduate Students

NAME	PERCENT_SUPPORTED	Discipline
Chia-Chieh Lin	0.10	
Xu He	0.20	
Zhuang Wu	0.10	
Wenwen Zhang	0.20	
Haoyang Fan	0.10	
FTE Equivalent:	**0.70**	
Total Number:	**5**	

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Wei-Yin Loh	0.10	
FTE Equivalent:	**0.10**	
Total Number:	**1**	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period:	0.00
The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:.....	0.00
The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:.....	0.00
Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):.....	0.00
Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:.....	0.00
The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense	0.00
The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields:	0.00

Names of Personnel receiving masters degrees

NAME	
Wenwen Zhang	
Haoyang Fan	
Total Number:	**2**

Names of personnel receiving PHDs

NAME	
Zhuang Wu	
Xu He	
Total Number:	**2**

Names of other research staff

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

(1) Handling of missing values and comparison of techniques.

One of the most difficult but important problems in statistics is how to construct a prediction model that can deal effectively with missing data values. This problem has two parts: when there are missing values in the data used to construct the model, and when there are no missing values in the training data, but missing values occur in the data used for future prediction. Many solutions have been proposed, but there are few comparative studies. During the period of this grant, an effective solution was found for use in the GUIDE classification and regression tree algorithm. If missing values occur in the training data, GUIDE provides a "missing" category to hold missing values when it performs contingency table chi-squared tests for split variable selection. Besides enabling GUIDE to use all the data for variable selection at each node of the tree, this also makes the algorithm sensitive to informative missingness, where data are not missing purely by chance. After a split variable is selected at a node, GUIDE considers three types of split on that variable for the node. The first type is to put all missing values in one branch of the split and all nonmissing values in the other branch. This allows GUIDE to detect missingness dependent on the response variable. The second type of split searches over the nonmissing data values to find the optimal split point, c, such that all nonmissing values less than c and all missing values are sent to the left branch. The third type is similar to the second type, except that all missing values are sent to the right branch instead of the left branch. If there are no missing values in the split variable, missing values in future data to be predicted are sent to the branch with the larger number of training cases. This strategy permits GUIDE to use all the data, all the time. Preliminary results from a large-scale empirical comparative study of GUIDE against other tree and non-tree methods indicate that this strategy, on average, better than other standard methods such as mean/mode imputation and complete cases. Details of the method for classification are reported in Loh (2009).

(2) Variable importance scores and thresholds for large p small n problems.

Data sets with more variables than observations have become increasingly frequent. Because many statistical methods require the number of observations to exceed the number of variables, there is an urgent need to find effective solutions for variable selection. For regression, the LASSO and similar solutions select variables by fitting a linear regression model to the data with a combination of L1 and L2 loss functions, which forces some variable coefficients to be zero and hence declares the others to be important. A major weakness of this approach is the assumption of an underlying regression model. Another weakness is that missing data values need to be estimated in advance. In the latter situation, the performance of the method depends critically on the missing value estimation method. During the period of this grant, an importance scoring method was implemented in the GUIDE algorithm. The key idea is to employ the chi-squared statistics that are already computed by GUIDE during model construction. By arguing that the chi-squared statistics are approximately mutually independent in the "null" case where all predictor variables are independent of the response variable, the statistics may be combined over the nodes of the tree to form an overall measure of importance for each variable. Furthermore, by approximating these scores with a single scaled chi-squared distribution, a threshold for separating the important from unimportant variables can be obtained. The availability of a threshold is a major accomplishment because such thresholds are not generally provided by previous importance scoring methods, which significantly limits their usefulness. The technique is first reported in Loh (2012).

(3) Multiresponse and longitudinal data.

Although there exist a few regression tree algorithms for longitudinal response variables, all of them are afflicted with variable selection bias, wherein variables that permit more splits are more likely to be selected, even when all the variables are independent of the response. Further, because these methods fit a linear mixed model to the data in each node, they are unstable and can be highly compute-intensive if the data set is large. During the period of this grant, a completely new approach to the problem was implemented in GUIDE. The key idea is to treat each longitudinal series as a random curve and then use the predictor variables to split the data by grouping the curves according to their shapes. As a result, no model assumptions are required and the observation time points and their number may be fixed or random. The same technique can be applied to multiresponse data where each subject is observed on two or more response variables. Details of the method are reported in Loh and Zheng (2103).

(4) Subgroup identification for censored and uncensored responses.

Difficult diseases such as cancer are hard to treat because not all patients respond equally to any given drug. As a result, the average efficacy of a new drug for all patients tends to be low, making it difficult to get approval from regulatory agencies. Current industry thinking is to identify subgroups of the patient population, defined in terms of characteristics such as gender, family history of disease, etc., as well as genetic traits, for which a drug has an enhanced effect. Because a regression tree naturally divides the sample and hence the population into subgroups of this type, there have been several attempts to use this approach to solve the problem. During this reporting period, the GUIDE algorithm was extended to provide a number of different alternative solutions, depending on whether the subgroups are to be defined in terms of prognostic or predictive factors. Results based on real and simulated data sets indicate that the GUIDE solutions are superior to the previous ones in terms of (i) accuracy in identification, (ii) computational speed, (iii) bias in selection of variables used to split the nodes, and (iv) extensibility to comparisons of more than two treatments and to censored response variables. A manuscript of the results is being prepared for publication.

Technology Transfer